COMPARISON OF DIFFERENT MULTIPLE IMPUTATION TECHNIQUES WITH COMPLETE CASE ANALYSIS IN A RANDOMIZED CONTROLLED TRIAL

Archan Chakraborty, under the guidance of Dr. B. Antonisamy,
Department of Biostatistics, Christian Medical College, Vellore, Tamil Nadu, India.

Introduction

- ❑ Randomized controlled trials (RCT) are considered the gold standard for evaluating a direct causal link between an intervention and outcome.
- ❑ A well-designed and conducted RCT provides an efficient and unbiased estimate of effect size when all observations required by the study protocol have been obtained. Difficulties can arise if some observations are missing.
- ❑ Missing observations reduce the effective sample size which leads to loss of statistical power and also the conclusions drawn from clinical trials with missing, which leads to bias.
- ❑ Generally the missing data will fall into any of the following pattern namely, Missing Completely At Random (**MCAR**),Missing At Random (**MAR**),Missing Not At Random (**MNAR**).
- ❑ In our data the type of all the missing values are MAR

Methodology

- ❑ In multiple imputation, missing values for any variable are predicted using existing values from other variables.
- ❑ This process is performed multiple times, producing, multiple imputed data sets (hence the term "multiple imputation (MI)").
- ❑ We have used three MI methods to impute the missing values. They are **Multivariate imputation by chained equations (MICE)**, **Likelihood Based Expectation Based Expectation Maximization (EM)** and **Random Forest (RF).**

MICE

•This chained equation process can be broken down into FIVE general steps:

•**Step 1:** A mean imputation, is performed for every missing value in the dataset.

•**Step 2:** The "place holder" mean imputations for one variable ("var") are set back to missing.

•**Step 3:** The observed values from the variable "var" in Step 2 are regressed on the other variables in the imputation model, "var" is the dependent variable in a regression model and all the other variables are independent variables in the regression model.

•**Step 4**: The missing values for "var" are then replaced with predictions (imputations) from the regression model. When "var" is subsequently used as an independent variable in the regression models for other variables, both the observed and these imputed values will be used.

•**Step 5:** Steps 2 through 4 are repeated for a number of cycles, with the imputations being updated at each cycle.

EM Algorithm

Handling missing data with Maximum Likelihood on all available data (takes that the missing values should be MAR or MCAR The EM algorithm uses a two-step iterative procedure where missing observations are filled in, or imputed, and unknown parameters are subsequently mated.

- In the first step (the E step), missing values are replaced with the conditional expectation of the missing data given the observed data and an initial estimate of the covariance matrix so-called FIML) is a very useful technique. The only assumption EM To illustrate, suppose a mean vector and covariance matrix, $\theta = (\mu, \Sigma)$ is sought for an n × K data matrix, Y, that contains sets of observed and missing values (Y_{obs} and Y_{mis} respectively). Using the observed values (Y_{obs}) and current parameter Estimates ($\theta^{(t)}$) the calculations for the sufficient statistics at the t^{th} iteration of the E step is

$$E\left(\sum_{i=1}^{n} y_{ij} \mid Y_{obs}, \theta^{(t)}\right) = \sum_{i=1}^{n} y_{ij}^{(t)}, \; j = 1, \dots, K$$

$$E\left(\sum_{i=1}^{n} y_{ij} y_{ik} \mid Y_{obs}, \theta^{(t)}\right) = \sum_{i=1}^{n} \left(y_{ij}^{(t)} y_{ik}^{(t)} + c_{ijk}^{(t)}\right), \; j, k = 1, \dots, K$$

where

$$y_{ij}^{(t)} = \begin{cases} y_{ij}, & \text{if } y_{ij} \text{ is observed} \\ E\left(y_{ij} \mid Y_{obs}, \theta^{(t)}\right), & \text{if } y_{ij} \text{ is missing} \end{cases}$$

And

$$c_{jkl}^{(t)} = \begin{cases} 0, & \text{if } y_{ij} \text{ or } y_{ik} \text{ is observed} \\ \mathrm{Cov}\left(y_{ij}, y_{ik} \mid y_{obs}, \theta^{(t)}\right), & \text{if } y_{ij} \text{ or } y_{ik} \text{ are missing} \end{cases}$$

Thus, missing values of y_{ij} are replaced with conditional means and covariances given the observed data and the current set of parameter estimates.

EM Algorithm

- ❑ In the second step (the M step), ML estimates of the mean vector and covariance matrix is obtained just as if there were no missing data using the sufficient statistics calculated at the previous E step. Thus, the M step is simply a complete-data ML estimation problem.
- ❑ The resulting covariance matrix and regression coefficients from the M step are then used to derive new estimates of the missing values at the next E step, and the process begins again

Random Forest

- ❑ Random forest (RF) imputation to be generally robust with performance improving with increasing correlation. It has advantages like **handling mixed types of missing data, address interactions and nonlinearity,** scale to high-dimensions while avoiding overfitting and yield measures of variable importance useful for variable selection.
- ❑ The miss forest algorithm data is imputed by regressing each variable in turn against all other variables and then predicting missing data for the dependent variable using the fitted forest with *p* variables. This means that p forests must be fit for each iteration.
- ❑ The *p* variables in the data set are randomly assigned into mutually exclusive groups of approximate size αp where $0 < \alpha < 1$. Each group in turn acts as the multivariate response to be regressed on the remaining variables (of approximate size $(1-\alpha)p$).

Over the multivariate responses, set imputed values back to missing. Grow a forest using composite multivariate splitting. Upon completion of the forest, the missing response values are imputed using prediction.

Objective

❑To review and compare the multiple imputation techniques in a randomised clinical trial data set with missing values, imputation and analysis.

Objectives:

1. To apply different multiple imputation methods in an RCT.
2. To evaluate if the prediction of the methods is similar compared to the complete case analysis.

Dataset Description

A double blinded, randomized controlled trial was performed to compare the effect of tamsulosin and placebo versus tamsulosin and tadalafil in lower urinary tract symptoms (LUTS) for the males above 45 years. A sample of **140 patient with LUTS** was included.

❑ Among them 71 people were given treatment 1 i.e, Tamsulosin + Placebo (Group 1) and 69 people were given treatment 2 i.e, Tamsulosin + Tadalafil (Group 1).

❑Among the follow up outcome values almost **17.1% values were missing** for outcome variables.

❑Among all of the 140 patients we have the complete dataset of 116 people.

❑For all the **primary outcome variables (IPSS total, voiding and storage)** and three **secondary outcome variables (IPSS QOL, Peak flow and PVRU)** there is no missing data in the baseline outcome variables. But in the follow up of all the primary outcome variables and all three secondary outcomes variables have missing data.

❑For these outcome variables in **treatment group 14(20.3%)** participants data are **missing** and in **placebo group 10(14.1%)** subjects data are **missing.**

❑For **IIEF5 in baseline, 7(10.1%)** participants have missing data in treatment group and **11(15.5%)** participants have missing data in placebo group. In the follow up, **18(26.1%)** participants data are missing in the **treatmen**t group and **20(28.2%)** participants data are missing in **placebo** group.

❑ The type of missing value is missing at random (MAR).

Results

Table 1: Baseline characteristics table

Variable	Tamsulosin and placebo (n=71)	Tamsulosin and Tadalafil (n=69)
Mean (SD)		
Age (years)	61.28(8.23)	58.87(8.16)
BMI (kg/m²)	24.39(3.38)	24.17(3.50)
Waist circumference (cm)	94.85(8.50)	95.17(7.75)
Co-morbidities, n (%)		
Diabetes Meletus	12 (16.9)	14(20.3)
Hypertension	18(25.4)	10(14.5)
Diabetes and hypertension	9(12.7)	6(8.2)
Others	6(8.4)	3(4.3)
None	26(36.6)	36(52.2)

Table 2: Complete Case Analysis Table

	Tamsulosin and Tadalafil Mean(SD)	Tamsulosin and Placebo Mean(SD)	Difference between means (95% CI)	p-value
Primary outcomes				
Total IPSS (3-months follow up)	9.16(3.69)	11.18(3.54)	-2.02 (-3.35 to -0.68)	0.003
Voiding (3-moths follow up)	5.18(2.90)	6.11(2.99)	-0.93 (-2.02 to 0.15)	0.091
Storage (3-months follow up)	3.98(1.95)	5.07(2.52)	-1.08 (-1.91 to -0.26)	0.01
Secondary outcomes				
IPSS Qol (3-months follow up)	2.16(1.17)	2.98(0.96)	-0.82 (-1.21 to -0.42)	<0.001
IIEF-5 (3-months follow up)	15.02(3.70)	10.75(3.36)	4.27 (2.88 to 5.66)	<0.001
Peakflow (3-months follow up)	13.20(3.91)	11.23(3.84)	1.97 (0.542 to 3.4)	0.007
PVRU (3-moths follow up) Median(IQR)	48 (28-66)	41(26-64)	1.93 (-11.11 to 14.96)	0.646

MICE Results

- ❑ We can see that, the mean is less in the treatment group compared to the placebo group for all the three primary outcome variables i.e., the difference in IPSS total (-2.07), void (-1.09) and storage (-1.01).
- ❑ From the p-values we can say that for IPSS total, IPSS storage, IPSS void shows there is significant difference between the means.
- ❑ For the secondary variables; IIEF5 score, IPSS QOL and Peakflow show there is significant difference between the means but PVRU shows there is no significant difference between the groups.

EM Results

- ❑ we can see that after 3 months the mean is less in the treatment group compared to the placebo group for all the three primary outcome variables i.e., the difference in IPSS total **(-1.62)**, void (**-0.82**) and storage (**-0.81**).
- ❑ From the p-values we can say that for IPSS total and IPSS storage shows there is significant difference between the means but IPSS void shows there is no significant difference between the means.
- ❑ For the secondary variables; IIEF5 score, IPSS QOL and Peak flow shows there is significant difference between the means but PVRU shows there is no significant difference between the two groups distribution.

Random Forest Results

- ❑ we can see that after 3 months the mean is less in the treatment group compared to the placebo group for all the three primary outcome variables i.e., the difference in IPSS total (**-1.65**), void (**-0.73**) and storage (**-0.92**).
- ❑ From the p-values we can say that for IPSS total and IPSS storage shows there is significant difference between the means but IPSS void shows there is no significant difference between the means.
- ❑ For the secondary variables; IIEF5 score, IPSS QOL and Peakflow shows there is significant difference between the means but PVRU shows there is no significant difference between the group distributions.

Conclusion

Based on the datasets, the following results are obtained,

- ❑ The combination of tamsulosin and tadalafil produced significantly better improvements in LUTS. In complete case (CC) analysis, except IPSS void (**p-value 0.091**) and PVRU (**0.646**) other outcomes shows significant difference in means between treatment and placebo groups.
- ❑ EM imputation (**for IPSS void p-value is 0.075**, **for PVRU p-value is 0.359**) and Random Forrest (**for IPSS void p-value is 0.114, for PVRU p-value is 0.521**), these two methods support the claim of CC analysis with their results. But for MICE except PVRU (**p-value 0.595**) all the other variables have shown significant difference in means between treatment and placebo groups.

References

- Little, R.J.A. and Rubin, D.B. (1987) Statistical Analysis with Missing Data. John Wiley & Sons, New York.
- Ramakrishnan V, Wang Z. Analysis of Data from Clinical Trials with Treatment Related Dropouts. Communications in Statistics – Simulation and Computation. 2005 Apr 1;34(2):343–53.
- Raghunathan, Trivellore & Lepkowski, James & Hoewyk, John & Solenberger, Peter(2000). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. Survey Methodology.